ScienceDirect

journal homepage: http://www.elsevier.com/locate/euprot

# Comparison of peptide and protein fractionation methods in proteomics

*Ekaterina Mostovenko*[a], *Chopie Hassan*[b], *Janine Rattke*[a], *André M. Deelder*[a], *Peter A. van Veelen*[b], *Magnus Palmblad*[a,*]

[a] *Biomolecular Mass Spectrometry Unit, Department of Parasitology, Leiden University Medical Center, PO Box 9600, 2300 RC Leiden, The Netherlands*

[b] *Immunohematology and Blood Transfusion, Leiden University Medical Center, PO Box 9600, 2300 RC Leiden, The Netherlands*

ARTICLE INFO

*Article history:*
Received 4 March 2013
Received in revised form
26 August 2013
Accepted 13 September 2013

*Keywords:*
Strong cation exchange chromatography
Isoelectric focusing
SDS-PAGE
Comparison
Taverna
Scientific workflows

ABSTRACT

Multiple fractionation or separation methods are often combined in proteomics to improve signal-to-noise and proteome coverage and to reduce interference between peptides in quantitative proteomics. Furthermore, a given fractionation method provides additional information on the analytes, such as molecular weight, hydrophobicity or isoelectric point that can be used to improve identification, and to discover protein splice variants or large post-translational modifications. Here we describe a Taverna scientific workflow for analysis and comparison between strong cation exchange (SCX) chromatography, peptide isoelectric focusing (pIEF) and SDS-PAGE performed using robust capillary LC and ion trap tandem mass spectrometry.



## 1. Introduction

Even with the recent improvement in speed and sensitivity of tandem mass spectrometry and performance of liquid chromatography systems, loading capacity and ion suppression still limit the coverage of complex samples, such as in proteomics. Thus, the prefractionation or reduction of complexity of samples is still beneficial in most analyses, when sufficient amounts of material are available. In general, each fraction contains a "simplified" mixture of peptides/proteins enabling identification and possibly quantitation of more peptides and proteins, including those of lower abundance. At the same time, fractionation adds information about the analytes without any additional analytical effort. This information can be used together with the tandem mass spectrometry data in the validation of peptide-spectrum matches.

A wide range of fractionation strategies for peptides and proteins are generally available, often combined in multidimensional methods or systems. Any type of chromatographic separation can be used at the protein level, including ion exchange [1], reversed phase [2], hydrophobic interaction

* *Corresponding author*. Tel.: +31 715269526.
E-mail address: n.m.palmblad@lumc.nl (M. Palmblad).

2212-9685 
http://dx.doi.org/10.1016/j.euprot.2013.09.001

[3] or size exclusion [4,5], prior to digestion. Ion exchange chromatography is frequently combined with reversed-phase chromatography, also at the peptide level, either off-line or on-line in the same column (MudPIT) [6]. Other popular methods include the Gelfree® fractionation system and SDS-PAGE [7]. The latter involves protein fractionation according to molecular weight, slicing the entire gel lane containing the proteins and then digesting the proteins in the gel. Isoelectric focusing of peptides or proteins can be done in capillaries [8,9], segmented tubes [10–12], gels [13] or liquid compartments connected by a gel [14].

In this work we have attempted to compare, with as little bias as possible, three very different and commonly used fractionation methods for two very different types of samples. We compared SDS-PAGE fractionation at the protein level [7], with Off-Gel™ isoelectric focusing, fractionating according to the isoelectric point [13], and strong cation exchange (SCX) chromatography, separating based on size and charge at a fixed pH [1], both at the peptide level.

Several previous studies have already been published for comparing these and other fractionation methods [15–17]. However, the choice of the best method likely also depends on the sample. We therefore compared the same three methods using exactly the same protocols for two different biological samples – an *Escherichia coli* whole cell lysate and human plasma. The *E. coli* cell lysates are easy to work with and not dominated by a few proteins. Human plasma on the other hand, is dominated by a small number of proteins, with albumin making up 45–50% of the total protein content, immunoglobulin G and transferrin another 8–20% and 3–7%, respectively [18]. The 20 most abundant proteins constitute more than 99% of the total protein content in plasma [18]. Both samples are easily obtained in large (even gram) quantities, making it possible to use almost any method for fractionation, from preparative scale chromatography to microfluidic methods coupled directly to the mass spectrometer.

The three compared methods each contribute information about a different peptide or protein property. This information can be used by some algorithms and pipelines to validate peptide and/or protein identification and remove erroneous identifications. In SDS-PAGE, the position of the protein on the gel has direct relationship with its molecular weight. When the measured protein molecular weight is compared with that predicted from the genome and used for the peptide identification, splicing events or post-translational processing could be detected. In IEF, the distribution of the peptides corresponds to their pI, which can also be predicted, albeit not with perfect accuracy. Finally, in SCX, the elution time (*i.e.* fraction number) depends on the size and charge of the peptides at the system pH [19], which may also be possible to predict from the peptide sequence. The Trans-Proteomic Pipeline (TPP) [20] already provides a standard score (also known as Z-score) for peptides based on their pI, and the same can in principle also be used for SCX chromatography. Indeed, the use of pI information to decrease the false discovery rate for IEF fractionated samples has been already demonstrated by other groups [21,22]. As part of the work presented here we also developed a general data analysis method and implemented this in a Taverna scientific workflow. The workflow compares multiple fractionation methods with respect to peptide and protein coverage while also extracting additional information on the peptides and proteins from each fractionation method. This information can be used for validation of peptide identifications and detection of splicing or post-translational events. We used this workflow to perform and visualize the comparison between the three different fractionation techniques for the two different types of samples, and briefly discuss the applicability of each method for each type of sample.

## 2. Materials and methods

For this study we compared three different separation approaches for two types of samples (human plasma and *E. coli*). Both groups of samples were treated similarly to enable comparison between methods to determine their suitability for different kinds of samples.

### 2.1. *Sample preparation*

Human plasma from healthy volunteers was collected into BD Vacutainer® tubes with 18.0 mg $K_2$:EDTA (K2E, REF 367525, BD Vacutainer Systems, Plymouth, UK) and immediately spun down at 1300 × *g* for 10 min at 21 °C then aliquoted and stored at −80 °C until use.

*E. coli* K12 strain MG1655 (ATCC® Number 47076, ATCC, Manassas, VA) was grown overnight in 4 × 25 mL Luria–Bertani (LB) medium in 50 mL Falcon tubes. The optical density at 600 nm (OD600) was 2.1. Then all cells were spun down and the supernatant removed. The pellets were resuspended in 10 mL warm (37 °C) PBS to pool all cells and gently spun down at 194 × *g* at 37 °C for 5 min. After the supernatant was removed, all pellets were rinsed with 1 mL PBS, transferred to a 1.5-mL Eppendorf tube and spun down again for 10 min at maximum speed (16,100 × *g*) at 4 °C. The wet pellet was weighed and 5 mL of the BugBuster® Master Mix (Novagen, Merck KGaA, Darmstadt, Germany) was added per gram cell paste. Cells were incubated at room temperature on a shaking platform at low speed for 20 min. After the insoluble cell debris was removed by centrifugation at 16,100 × *g* for 20 min at 4 °C, the supernatant was stored at −80 °C until used.

### 2.2. *In-solution digestion*

Two mg of each sample were digested using trypsin. To each sample DTT in 25 mM ammonium bicarbonate (ABC) was added to its final concentration 10 mM and incubated for 45 min at 56 °C to reduce cystines. After alkylation for 1 h at room temperature with 25 mM iodoacetomide also in 25 mM ABC trypsin (sequencing grade, Promega, Madison, WI) was added in the ratio 1:100 (trypsin:sample) and kept for 10 h at 37 °C. Digestion was quenched with 10% TFA with the final concentration of TFA 0.1–1.0%. Resulting samples were desalted using Oasis HLB cartridges and aliquoted in 100 and 200 μg for IEF and SCX, respectively.

### 2.3. *Desalting and solid phase extraction*

Prior to fractionation both samples were desalted using Oasis HLB cartridges (Waters, Milford, MA). Cartridges were first

activated with methanol and equilibrated with 50% acetonitrile (ACN) in water according to the manufacturer's protocol. The sample was applied and washed 4 times with 500 μL water. The peptides were eluted into a fresh Eppendorf tube with 800 μL 50% ACN.

Fractions collected after the separation were desalted with solid-phase extraction (SPE) using C18 OMIX tips (Agilent Technologies, Waldbronn, Germany). Tips were first wetted with 50% ACN in water, washed and equilibrated with water containing 0.1% TFA. Samples were acidified with TFA, applied onto tip, washed again and then eluted with 50 μL 50% aqueous ACN containing 0.1% TFA. Acetonitrile was evaporated after each cleaning step.

### 2.4. *Strong cation exchange*

SCX was performed on a Dionex UltiMate 3000 (Thermo Fischer Scientific, Waltham, MA) at a flow rate of 200 μL/min. Tryptic peptides (200 μg) were loaded onto a 100 mm × 2.1 mm PolySULFOETHYL A™ (PolyLC, Columbia, MD) column with 3 μm packing material and eluted with a linear gradient using ACN/potassium phosphate buffers (buffer A – 20% ACN/80% 10 mM potassium phosphate, pH 2.9; buffer B – 20% ACN/80% 10 mM potassium phosphate, 500 mM potassium chloride, pH 2.9). The elution programme was 100% buffer A for 10 min, continued by a short (1 min) gradient of 0–3% of buffer B, followed by a gradient of 3–15% for 19 min, a 15–45% gradient for 15 min and a 45–100% gradient for 2 min. At the end of the gradient the column was kept at 100% buffer B for 7 min and then for 10 min in buffer A. Flow-through fractions (48 in total) were collected into a 96-well plate from 5 to 55 min. Adjacent fractions were combined pairwise to obtain 24 fractions and then desalted with SPE (described above).

### 2.5. *Isoelectric focusing*

For peptide IEF separations, the Off-Gel Agilent 3100 fractionator (Agilent Technologies) was used. A modified method was applied by addition of 1 M urea to the buffer sample and rehydration buffer, instead of 5% glycerol only. Tryptically digested and desalted peptides (100 μg in total) were resuspended in a modified IPG buffer that contained 1 M urea in addition to the 3–10 pH linear IPG buffer (GE Healthcare, Uppsala, Sweden). Sample volumes of 150 μL/well were loaded onto a commercially available 24-cm IPG strips with a linear 3–10 pH gradient (GE Healthcare) after rehydration of the gel for 20 min in 40 μL/well rehydration solution. Cover fluid (mineral oil, Agilent Technologies) was applied to both ends of the gel strip. The focusing method OG24PE01, as supplied by the manufacturer, was used for 24-well fractionations. Fractions were recovered in separate Eppendorf tubes, cleaned by SPE as described above and store at −80 °C till use.

### 2.6. *SDS-PAGE and in-gel digestion*

Protein concentration was measured by the bicinchoninic acid (BCA) protein assay kit (Thermo Fischer Scientific) and 30 μg of proteins per sample was loaded on a 1-mm 10-well 4–12% NuPAGE® Bis-Tris gel (Invitrogen, Carlsbad, CA). Proteins were separated in the gel for 1 h at 180 V, after which the gel was stained in NuPAGE® Colloidal Blue (Invitrogen) overnight at room temperature and destained with milli-Q water until the background was transparent.

The gel lane with separated proteins was cut into 48 identical 1.5-mm × 5-mm slices using a MEE1.5-5-48 disposable gel cutter (Gel Company Inc., San Francisco, CA). Each gel piece was placed into one well in a 96-well polypropylene PCR plate (Greiner Bio-One, Frickenhausen Germany). Destaining of the gel pieces, DTT reduction and IAA alkylation were performed according to the previously published protocol [23]. In-gel tryptic digestion was performed in 30 μL of 25 mM ABC containing 5 ng/μL trypsin (sequencing grade, Promega, Madison, WI) for 6 h at 37 °C. The resulting peptides were TFA-extracted according to the previously described protocol [23]. The extracts were pooled pairwise to obtain 24 total fractions as for SCX.

### 2.7. *LC–MS/MS analysis*

Prior to LC–MS/MS analysis all samples were dried down and reconstituted in 25 μL 0.1% TFA. The analysis was performed using a splitless NanoLC-Ultra 2D plus (Eksigent, Dublin, CA) for parallel ultra-high pressure liquid chromatography (UHPLC) with an additional loading pump for fast sample loading and desalting. The UHPLC system was configured with 300 μm-i.d. 5-mm PepMap C18 trap columns (Thermo Fischer Scientific) and 15-cm 300 μm-i.d. ChromXP C18 columns (Eksigent). Ten microliter (full loop) of each fraction were injected onto the column and separated by a 45-min linear gradient from 4 to 33% acetonitrile in 0.05% formic acid with 4 μL/min flow rate. The UHPLC system was coupled on-line to an amaZon ETD speed high-capacity 3D ion trap with CaptiveSpray source (Bruker Daltonics, Bremen, Germany). After each MS scan, up to ten abundant multiply charged species in the *m*/*z* 300–1300 range were automatically selected for MS/MS but excluded for 1 min after having been selected twice. The UHPLC system was controlled using HyStar 3.4 with a plug-in from Eksigent and the amaZon ion trap by trapControl 7.0, all from Bruker.

### 2.8. *Data analysis*

All acquired tandem mass spectrometry data were processed in one batch using the Taverna workbench [24]. Taverna can invoke a number of services, including local Java Beanshell scripts, R (using an R server) and a wide range of Web services, enabling combination of sequence database search, analysis and visualization in a single workflow. Built-in tools for parsing XML-files simplify information retrieval and large datasets can be remotely processed on a grid or cloud using the Taverna Engine [25]. The workflow used here converts raw data to mzXML [26] using compassXport 3.0 (Bruker) and passes this, along with the sequence database and search parameters to X!Tandem [20,27] in the TPP [20]. The X!Tandem scores are converted to pepXML [20], modelled and converted to probabilities for each peptide-spectrum match by PeptideProphet [28]. The X!Tandem search was here done against the UniProt human reference proteome set (2012_02, canonical sequences only) and the UniProt *E. coli* reference set (2010_01) with the monoisotopic mass error (±0.5 Da), carbamidomethylation as fixed modification, the *k*-score plug-in [20] and allowing for

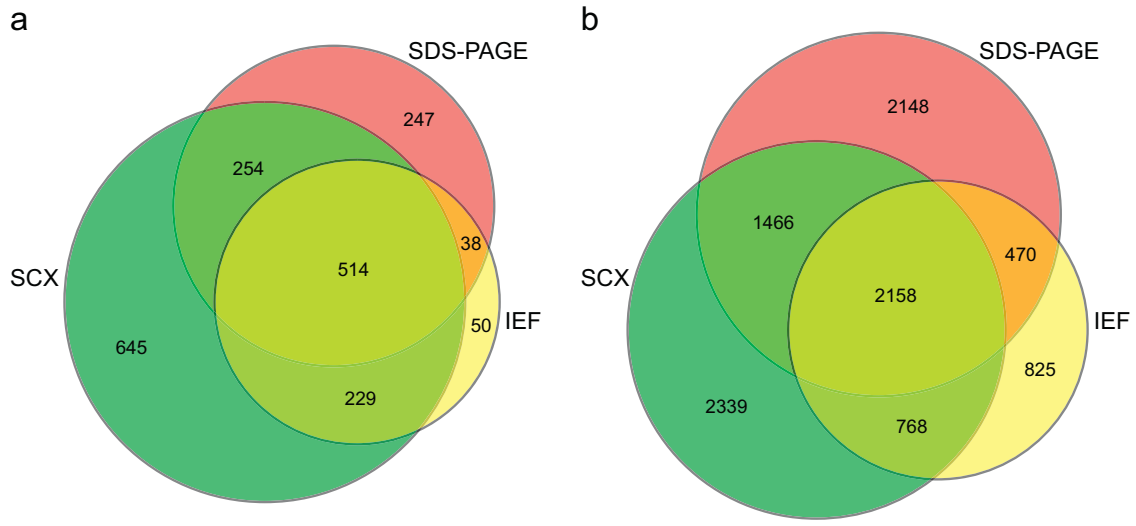


**Fig. 1 – Comparison of the number of unique peptide identifications from SDS-PAGE, SCX and IEF datasets for human plasma (a) and *E. coli* (b). The total numbers of unique peptides identified in human plasma were 1053 with SDS-PAGE, 1642 with SCX and 831 with IEF. In the *E. coli* samples, 6242 unique peptides were identified with SDS-PAGE, 6731 with SCX and 4221 with IEF, respectively. The numbers of unique *proteins* identified from the *E. coli* samples were relatively similar between the three methods: 1037 (SDS-PAGE), 942 (IEF) and 1139 (SCX), all with 1% FDR as estimated by ProteinProphet. For human plasma, the numbers were much smaller as expected, with 126 (SDS-PAGE), 128 (IEF) and 183 (SCX) unique proteins identified with 1% ProteinProphet-estimated FDR.**

isotope error. After PeptideProphet analysis, the resulting lists of peptide/protein identifications with 0.95 probability cut-off (<1% FDR) were analyzed and compared in the Rshell script in the same workflow. For each peptide within one IEF fraction, pI values predicted by attached function in TPP (based on pK values from Bjellqvist et al. [29]) were extracted and the pI Z-scores were calculated as a distance in standard deviations from the mean. To compare the pI of true and false matches, a search was also done against a decoy database generated by randomizing the *E. coli* database with make_random (http://www.ms-utils.org/make_random.html). For SDS-PAGE, the protein molecular weight was calculated from the sequences downloaded from the UniProt website directly in Taverna workflow as these are not kept in the pepXML results. The entire processing workflow is available in myExperiment (http://www.myexperiment.org/workflows/3486.html).

## 3. Results

In this work we compared SDS-PAGE, SCX and IEF separation strategies for two different types of samples with a robust analysis method. For both samples the highest proteome coverage we observed with SCX (Fig. 1) identifying 1139 proteins from 6731 peptides in the *E. coli* sample, and 183 proteins from 1642 unique peptides in crude plasma. In the recent work of Hassan et al. [30], SCX was also demonstrated to be better than IEF, as measured by the number of identified peptides. When comparing the number of identified proteins, SDS-PAGE gave the lowest coverage for plasma, similarly to a previous comparison using HeLa cells [31]. The mean protein sequence coverage followed a similar pattern to the peptide and protein identifications with the highest coverage in the SCX fractions (35% for the *E. coli* digest and 29% for the human plasma) than in SDS-PAGE (34% and 23% respectively) and IEF (27% and 24%). These numbers are relative to the full UniProt FASTA sequences. The actual sequence coverages are slightly higher, if the parts of the sequences that are removed in the mature proteins are also subtracted from the denominators in the calculation of sequence coverages.

To define the quality of the separation we looked at the distribution of the number of peptides identified per fraction (Fig. 2). When separated with SCX, most peptides were found in one fraction. In IEF, the majority of peptides is also identified in one fraction, however the number of peptides found in two or more fractions is much higher compared to SCX. Pie charts in Fig. 2 illustrate the peptide distribution for human plasma sample. For *E. coli* the observation is consistent (data not shown).

A further motivation behind this study was to produce, from the same samples, similar datasets using the three different peptide and protein fractionation techniques to illustrate the value of the additional information on the analytes that can be automatically obtained from a particular method. Using a pH indicator, we observed that the peptides in IEF separate more or less linearly in the pH gradient independent of the nature of the sample (Fig. 3a, top). Thus the calculated pI can be plotted against the actual fraction number and possible outliers would most likely be false identifications (Fig. 3a, bottom). The predicted pI appear to change in more discrete steps compared to the smooth transitions of the pI indicator. Another way to represent this information is to calculate pI Z-score and visualize their distribution for each fraction separately using a histogram or a box-plot (Fig. 3b). The decoys have a wide distribution in Z-score (the unit determined by the standard deviation in predicted pI of the matches from the

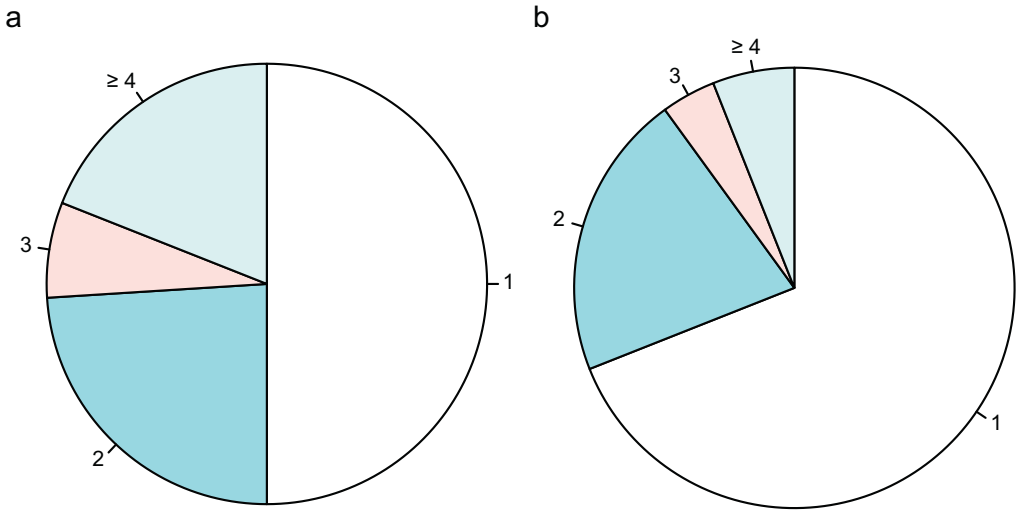


**Fig. 2 – Pie charts illustrating the percentage of peptides identified in one or more fractions after separation of human plasma by IEF (a) or SCX (b). During the SCX chromatography 48 fractions from 65 min gradient were collected and every two consecutive ones pooled together.**

correct database) and as expected with bias towards higher pI for fractions of low pI and towards lower pI for fractions of high pI, whereas the correct identifications are focused near the average pI of all peptides identified in the fraction.

SDS-PAGE, on the other hand, provides direct information about the proteins rather than the peptides. Predicted, based on the sequence, protein molecular weight plotted against its location on gel (fraction number) shows a clear correlation (Fig. 4). However, a number of outliers can still be identified for closer examination or discarding as false discoveries. As an example, the 20 kDa Alkyl hydroperoxide reductase subunit C (UniProt accession number P0AE08), in native conditions disulfide-linked homodimer, was observed at ~40 kDa (Fig. 4, arrow).

## 4. Discussion

The wide range of available fractionation techniques makes it challenging to choose the one best suited for a particular sample or biological research question. We performed in this work comparison of the described above techniques at the level of proteins (SDS-PAGE) and peptides (IEF and SCX). The major challenge in setting up such a study is to make the comparison "fair", given the differences in scale and practical implementation of the techniques, *i.e.* sensitivity levels, system volumes/flow rates and fraction collection. It is especially difficult to use the same amount of starting material for each method without diluting the sample or overloading one or more of the systems. In the case of SDS-PAGE, the maximum amount of protein that can be applied on the standard, commercially available, gel without overloading is around 30 μg. For the IEF and SCX methods, the equivalent amount of peptides would be too low due to the minimal volumes involved. The work of Hubner et al. [31] demonstrated that the best separation with IEF could be achieved with 50 μg material, while the maximum number of protein identifications was achieved with 250 μg. From our experience, the optimal condition for Off-Gel IEF (balance between good separation and wide proteome coverage) has been obtained when loading 100 μg. Even though SCX gave reasonable separation when loaded 100 μg of material, the system was far from its maximum loading capacity, leading us to increase the amount of proteins injected to limit sample dilution. For this reason we compromised and loaded different amounts to allow each fractionation technique to operate near its maximum capacity, taking into account the significant dilution in the IEF and SCX as compared to SDS-PAGE. Robustness and stability of the liquid-chromatography–mass spectrometry analysis is also important for the method comparisons in absence of internal standards or labels. To balance sensitivity and robustness, the choice was made to use the new CaptiveSpray (Bruker) source, accommodating higher flow rates and therefore more robust chromatography than the more sensitive but less stable nanoelectrospray.

One would expect SDS-PAGE to be a good choice for samples dominated by a small number of abundant proteins, such as plasma, as these abundant proteins can be confined to a few bands or fractions. In contrast, when performing the fractionation at the peptide level, peptides from the abundant proteins will be present in most if not all fractions. However, in this comparison, we demonstrated that SCX was clearly better in both peptide and protein yield. This proves that the loading capacity can be more important than the separation method or whether the fractionation is done at the protein or peptide level. The IEF approach gave the smallest number of identifications and showed the largest overlap with the other two techniques for both samples. Even though the work of Hubner et al. [31] showed that Off-Gel IEF gives higher number of protein identifications compared to SDS-PAGE in human cell lines, other recent work comparing SDS-PAGE, SCX, IEF and organelle fractionation have showed the opposite [32]. For the IEF system it is known that near the edges of the gel (at pH 3 and pH 10) it is common to see diffuse bands if the

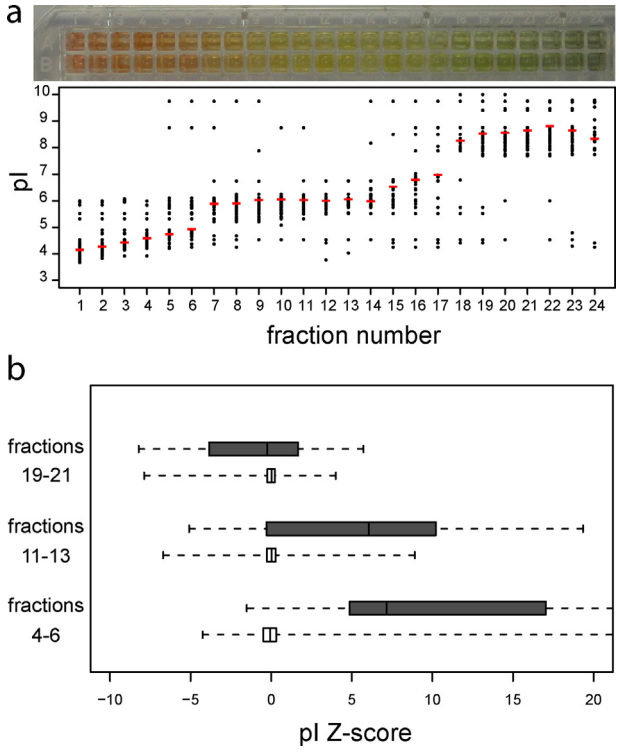


**Fig. 3 – Fractionation of human plasma and *E. coli* peptides by isoelectric focusing. The photograph shows the pH indicator from pH ~ 3 to pH ~ 10 in the fractions of plasma (top) and *E. coli* (bottom). The pH gradient appears reproducible and independent from the sample. The fraction yielding the largest number of spectrum matches for a peptide can be plotted against the predicted pI for the peptide (here showing only the IEF fractions in *E. coli*). The mean pI of the peptides in each fraction is marked with red bars. The pI information can be used to weed out false identifications. The box plot (b) illustrates the distribution of pI Z-scores with putatively correct matches in white and decoy matches in grey.**

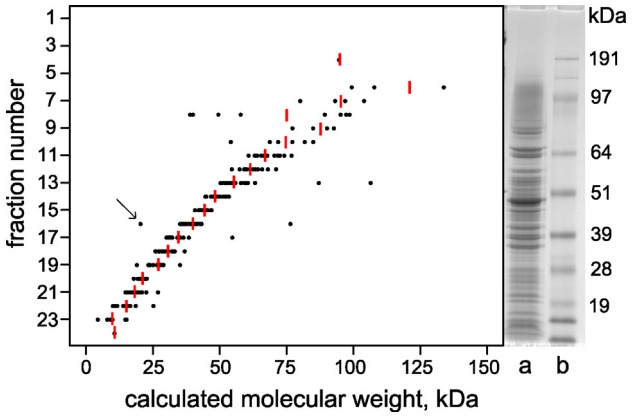


**Fig. 4 – Protein molecular weight distribution against fraction number in the *E. coli* sample (left). Median molecular weight per fraction is marked with vertical bars. SDS-PAGE of 30 μg of *E. coli* sample (a) and molecular weight marker (b) on the right. Alkyl hydroperoxide reductase subunit C (UniProt accession number P0AE08) is marked with an arrow.**

gel is stained, indicating less sharp separation. Consequently, peptides can be found in more than one fraction near the edges, increasing the redundancy of the data and reducing the number of different peptides that can be identified. During the SCX chromatography, fractions were collected every 60 s and subsequent fractions were pooled for the analysis to keep the number of fractions similar to those obtained with IEF and mass spectrometry analysis time constant throughout the whole experiment. The number of collected fractions with SCX is determined by the fraction collection method, which can easily be adjusted to any number, as long as the vials or wells can hold the volume and there are enough physical vials or wells in the fraction collector. Similarly for SDS-PAGE, any reasonable number of pieces can be cut, as long as the slices are not too thin to handle in a practical manner. Generally, more fractions lead to wider proteome coverage if the mass spectrometry time per fraction is constant. As the numbers of SCX fractions and gel slices are easy to vary, the defining factor for the number of collected fractions was the IEF system. For SDS-PAGE, we used an already existing and commercially available cutter enabling slicing the gel into 48 equal slices at once. Similarly, we used an existing method for collecting 48 SCX fraction and then pool the adjacent ones to obtain 24 total fractions for each separation method. Most peptides identified with SCX were found in a single fraction, showing that peptides elute in narrow peaks (maximum 2 min in a 65-min gradient). Compared to the studies conducted by Slebos et al. [17] and Elschenbroich et al. [33] demonstrating that IEF is superior to SCX in resolution, we used longer and better analytical column for SCX, with smaller bead size. Not surprisingly, in our experimental set-up, SCX had better resolution than IEF, defined as peptide overlap between fractions. The fractionation settings and the design of the comparison have more influence on the result than the nature of the sample.

For any scheme that uses information from the fractionation prior to the chromatographic separation on-line with the tandem mass spectrometer it is crucial that this information is preserved throughout the data analysis. This is easily accomplished by a systematic naming of files or by loading fractions in sequence into a microtiter plate. From each dataset, specific protein or peptide information could be extracted and used for filtering out spurious identifications. The theoretical model used for pI calculation is based on the peptide sequence and does not take influences of nearby residues into account, leading to a discrete rather than smooth distribution of pI in the IEF-separated samples. However, this information is used in the calculation of pI Z-scores for each peptide-spectrum match, assuming they derive from a fraction with a narrow pI distribution, and is already implemented in the TPP. Random, false (decoy) peptide-spectrum-matches can derive from peptides of any pI therefore having a wide span, whereas the correct identifications are concentrated around 0. For a perfect Gaussian distribution, the lower and upper quartiles, *i.e.* the "box", would be between Z-score −0.68 and 0.68. In the pI box plots in Fig. 3b, the lower and upper quartiles of the

putatively correct peptide identifications span a slightly smaller interval. This is likely due to a number of outliers caused by very abundant peptides being identified in many fractions and differences between calculated and real (experimental) pI.

Although some success has been reported in the prediction of peptide retention times in SCX [34,35], this has so far only been achieved with machine-learning techniques such as artificial neural networks, requiring tens or hundreds of thousands of peptide identifications to train the model. This makes the approach feasible only when very large collections of datasets are available. A simpler model could be plugged into the workflow as available on myExperiment. Since both SCX and IEF are primarily based on charge (SCX on the charge at a particular pH) it may be tempting to use a similar model for SCX prediction as for pI prediction in IEF. However, for the datasets used in this work, this did not produce a useful model.

Protein information derived from SDS-PAGE can indirectly indicate whether peptide identifications correspond to a protein that is likely to be present in the fraction from which the spectrum was acquired. However, as there are many reasons why the calculated and measured protein molecular weights may differ significantly, it is probably more sensible to use the protein level information to learn something about the proteins. Proteins located far above a curve fitted to the predicted molecular weights are larger than predicted (Fig. 4), which might be due to an incomplete sequence in the database, a large post-translational modification or a covalent protein complex. Hits below the curve indicate that the observed protein is only part of the predicted (database) protein sequence. If both explanations are implausible and the number of confident peptide-spectrum matches for a protein is small (given the total number of spectra acquired), the protein identification is likely incorrect. This assumption is supported by the relatively low probabilities for the peptide-spectrum matches for these proteins. In prokaryotic organisms, there is little post-transcriptional activity, such as splicing, that leads to multiple protein isoforms from the same gene or entry in the searched FASTA file. There are also fewer post-translational events decorating proteins with adducts large enough to be noticeable by SDS-PAGE. Therefore, the outliers are most likely false identifications, and their number is very small compared to those in eukaryotic samples. A few exceptions, such as covalent complexes, can still be identified though, as shown in Fig. 4. Using the SDS-PAGE information, false positives can be weeded out at the protein – rather than the peptide level, without influencing the probabilities assigned to the peptide-spectrum matches.

## 5. Conclusions

In shotgun proteomics, good coverage of complex samples still requires more than one dimension of fractionation or separation. However, not all separation methods are equally suitable for all types of samples and research questions. Here we compared three of the most commonly used techniques, SDS-PAGE, SCX and IEF, for two different and "typical" samples. The fractionation methods are based on different physicochemical properties and were performed at different levels – at the protein level with SDS-PAGE and at the peptide level with SCX and IEF. When comparing such different separation techniques, it is difficult to make a "fair" comparison. We kept the final number of fractions collected equal and the total mass spectrometry analysis time constant, but decided to compromise on the amount of protein used, performing the fractionation near the optimal conditions/highest capacity of each method. The number of collected fractions and MS instrument time were kept the same for the comparisons, even though the SDS-PAGE and SCX would likely have performed better if more fractions had been collected. Under the studied conditions, IEF showed the lowest coverage for both samples, which may be partly due to the dilution occurring during the run but also to suboptimal number of fractions in IEF. The extracted pI information gives an easily implemented method to filter out false peptide-spectrum matches. The SDS-PAGE approach resulted in better coverage of the proteome, while also providing molecular weight information on the proteins. We compromised the resolving power of the gel by pooling consecutive pairs of gel slices to keep the total number of fractions the same as for the IEF. There is no strict reason to believe that combining adjacent fractions is the most optimal way to reduce the number of fractions. By pooling two neighbouring fractions where most likely similar proteins are dominant, there will be suppression of the less represented ones. It is possible that it would be better to combine gel slices containing large and small proteins, even though the results would be more difficult to interpret manually.

Strong cation exchange provided the best sequence coverage as well as the largest number of peptide identifications (especially for *E. coli*) and protein identifications (particularly for plasma). The information of SCX retention times which could be used to improve sensitivity and lower the false discovery rate was not implemented. Although SCX is a very efficient separation technique for peptides and orthogonal to reversed-phase, it is most likely that it was the larger amount of sample that could be loaded on the SCX column, compared to the SDS-PAGE and IEF that contributed the most to the higher number of identifications.

The data analysis, from raw data to the graphs as they appear in the paper, could be performed entirely within one Taverna workflow, facilitating sharing not only raw data but also executable workflows. This allows other researchers to reproduce the analysis while varying input parameters or apply the same analysis workflow on their own data. Additionally, separate components of the workflow could be reused in different analyses or adopted for other tasks. The workflow executed local commands and took the data through the Trans-Proteomic Pipeline interfacing data analysis of three separate datasets in parallel using one parameter and one FASTA file piped to different processes assuring exactly the same conditions for each search. This workflow also fetched information from on-line databases, performed statistical analyses in R and plotted the results.

## Acknowledgements

The authors wish to thank Hans Dalebout and Oleg Klychnikov for technical assistance and useful discussions, and Yassene Mohammed for help with the Taverna workflows.

## REFERENCES


[1] Choudhary G, Horvath C. Ion-exchange chromatography. Methods Enzymol 1996;270:47–82.
[2] Howard GA, Martin AJP. The separation of the C-12–C-18 fatty acids by reversed-phase partition chromatography. Biochem J 1950;46:532–8.
[3] Hjerten S. Some general aspects of hydrophobic interaction chromatography. J Chromatogr 1973;87:325–31.
[4] Lathe GH, Ruthven CR. The separation of substances on the basis of their molecular weights, using columns of starch and water. Biochem J 1955;60:xxxiv.
[5] Porath J, Flodin P. Gel filtration: a method for desalting and group separation. Nature 1959;183:1657–9.
[6] Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, et al. Direct analysis of protein complexes using mass spectrometry. Nat Biotechnol 1999;17:676–82.
[7] Shapiro AL, Vinuela E, Maizel Jr JV. Molecular weight estimation of polypeptide chains by electrophoresis in SDS-polyacrylamide gels. Biochem Biophys Res Commun 1967;28:815–20.
[8] Chen J, Lee CS, Shen Y, Smith RD, Baehrecke EH. Integration of capillary isoelectric focusing with capillary reversed-phase liquid chromatography for two-dimensional proteomics separation. Electrophoresis 2002;23: 3143–8.
[9] Chen J, Balgley BM, DeVoe DL, Lee CS. Capillary isoelectric focusing-based multidimensional concentration/separation platform for proteome analysis. Anal Chem 2003;75: 3145–52.
[10] Hjerten S. Free zone electrophoresis. Chromatogr Rev 1967;9:122–219.
[11] Margolis J, Corthals G, Horvath ZS. Preparative reflux electrophoresis. Electrophoresis 1995;16:98–100.
[12] Horvath ZS, Gooley AA, Wrigley CW, Margolis J, Williams KL. Preparative affinity membrane electrophoresis. Electrophoresis 1996;17:224–6.
[13] Michel PE, Reymond F, Arnaud IL, Josserand J, Girault HH, Rossier JS. Protein fractionation in a multicompartment device using Off-Gel (TM) isoelectric focusing. Electrophoresis 2003;24:3–11.
[14] Xiao Z, Conrads TP, Lucas DA, Janini GM, Schaefer CF, Buetow KH, et al. Direct ampholyte-free liquid-phase isoelectric peptide focusing: application to the human serum proteome. Electrophoresis 2004;25:128–33.
[15] Gan CS, Reardon KF, Wright PC. Comparison of protein and peptide prefractionation methods for the shotgun proteomic analysis of *Synechocystis* sp. PCC 6803. Proteomics 2005;5:2468–78.
[16] Essader AS, Cargile BJ, Bundy JL, Stephenson Jr JL. A comparison of immobilized pH gradient isoelectric focusing and strong-cation-exchange chromatography as a first dimension in shotgun proteomics. Proteomics 2005;5:24–34.
[17] Slebos RJ, Brock JW, Winters NF, Stuart SR, Martinez MA, Li M, et al. Evaluation of strong cation exchange versus isoelectric focusing of peptides for multidimensional liquid chromatography–tandem mass spectrometry. J Proteome Res 2008;7:5286–94.
[18] Putnam FW. Alpha, beta, gamma, omega – the roster of the plasma proteins. In: Putnam FW, editor. The plasma proteins. 2nd ed. NY: Academic Press; 1975. p. 57–130.
[19] Peng JM, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC–MS/MS) for large-scale protein analysis: the yeast proteome. J Proteome Res 2003;2:43–50.
[20] Keller A, Eng J, Zhang N, Li XJ, Aebersold R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol 2005;1:1–8.
[21] Cargile BJ, Bundy JL, Freeman TW, Stephenson JL. Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. J Proteome Res 2004;3:112–9.
[22] Krijgsveld J, Gauci S, Dormeyer W, Heck AJ. In-gel isoelectric focusing of peptides as a tool for improved protein identification. J Proteome Res 2006;5:1721–30.
[23] Mostovenko E, Deelder AM, Palmblad M. Protein expression dynamics during *Escherichia coli* glucose-lactose diauxie. BMC Microbiol 2011;11:126.
[24] Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 2004;20:3045–54.
[25] Mohammed Y, Mostovenko E, Henneman AA, Marissen RJ, Deelder AM, Palmblad M. Cloud parallel processing of tandem mass spectrometry based proteomics data. J Proteome Res 2012;11:5101–8.
[26] Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, et al. A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol 2004;22:1459–66.
[27] Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 2004;20:1466–7.
[28] Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002;74:5383–92.
[29] Bjellqvist B, Hughes GJ, Pasquali C, Paquet N, Ravier F, Sanchez JC, et al. The focusing positions of polypeptides in immobilized Ph gradients can be predicted from their amino-acid-sequences. Electrophoresis 1993;14:1023–31.
[30] Hassan C, Kester MG, Ru AH, Hombrink P, Drijfhout JW, Nijveen H, et al. The human leukocyte antigen-presented ligandome of B lymphocytes. Mol Cell Proteomics 2013;12(7):1829–43.
[31] Hubner NC, Ren S, Mann M. Peptide separation with immobilized pI strips is an attractive alternative to in-gel protein digestion for proteome analysis. Proteomics 2008;8:4862–72.
[32] Antberg L, Cifani P, Sandin M, Levander F, James P. Critical comparison of multidimensional separation methods for increasing protein expression coverage. J Proteome Res 2012;11:2644–52.
[33] Elschenbroich S, Ignatchenko V, Sharma P, Schmitt-Ulms G, Gramolini AO, Kislinger T. Peptide separations by on-line MudPIT compared to isoelectric focusing in an off-gel format: application to a membrane-enriched fraction from C2C12 mouse skeletal muscle cells. J Proteome Res 2009;8:4860–9.
[34] Alpert AJ, Petritis K, Kangas L, Smith RD, Mechtler K, Mitulovic G, et al. Peptide orientation affects selectivity in ion-exchange chromatography. Anal Chem 2010;82:5253–9.
[35] Petritis K, Kangas L, Jaitly N, Monroe ME, Lopez-Ferrer D, Maxwell RA, et al. Strong cation exchange LC peptide retention time prediction and its application in proteomics. In: American Society for Mass Spectrometry (ASMS). 2008.